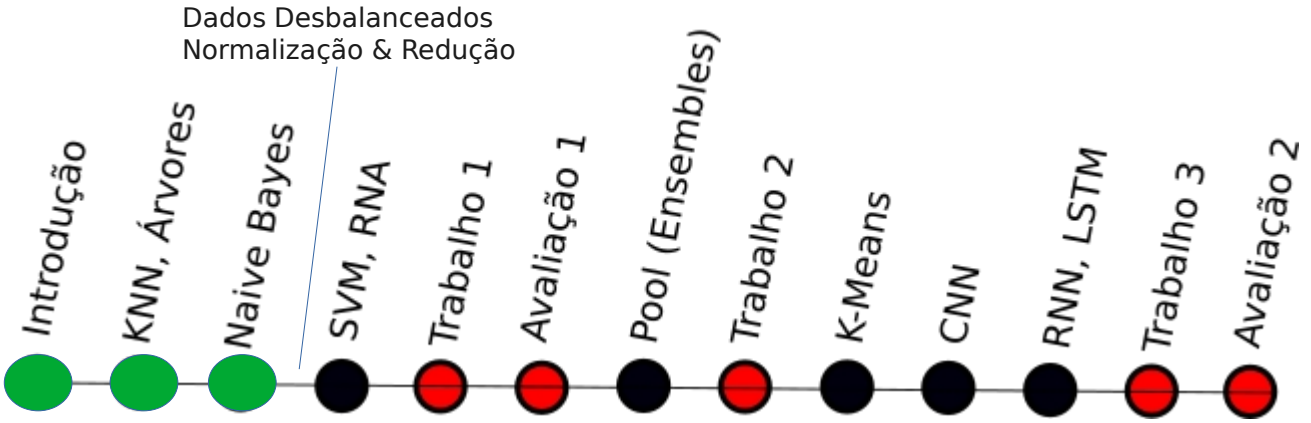
Normalização e Redução dos Dados

Prof. André Gustavo Hochuli

gustavo.hochuli@pucpr.br
aghochuli@ppgia.pucpr.br
github.com/andrehochuli/teaching

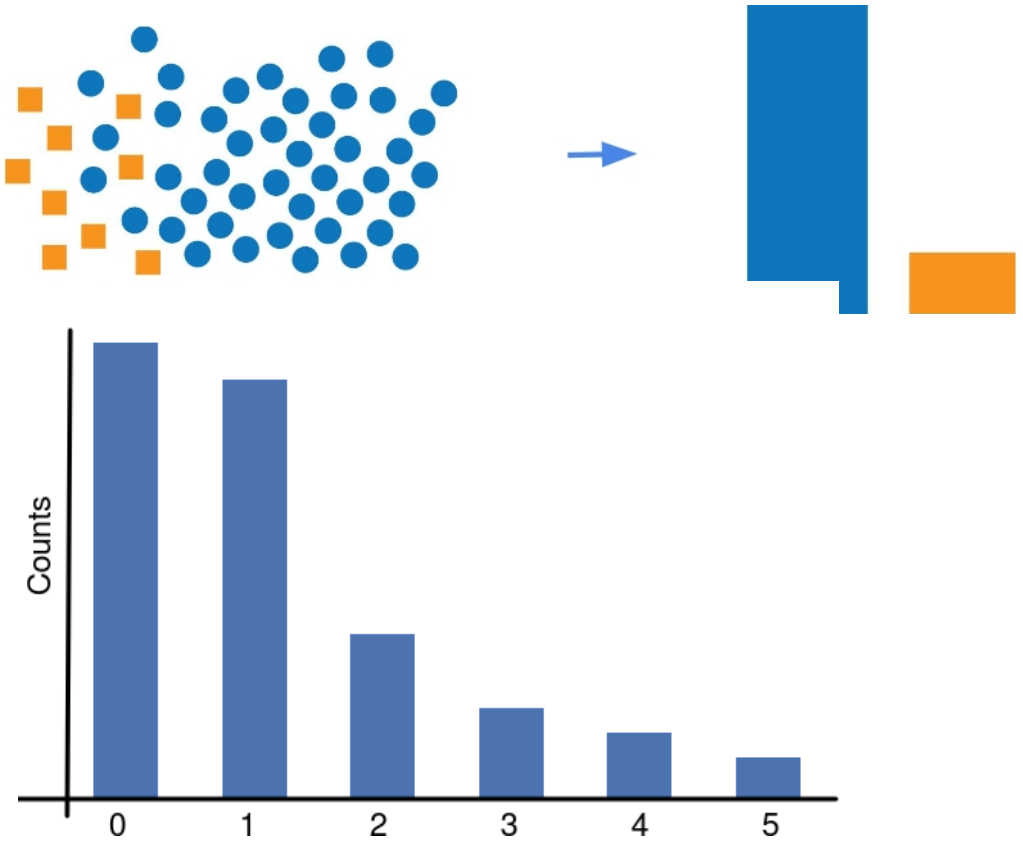
Plano de Aula

- Discussões Iniciais
- Normalização
- Redução
- Visualização
- Exercícios



Discussões Iniciais

- Dados Desbalanceados
  - KNN, NB e Árvores



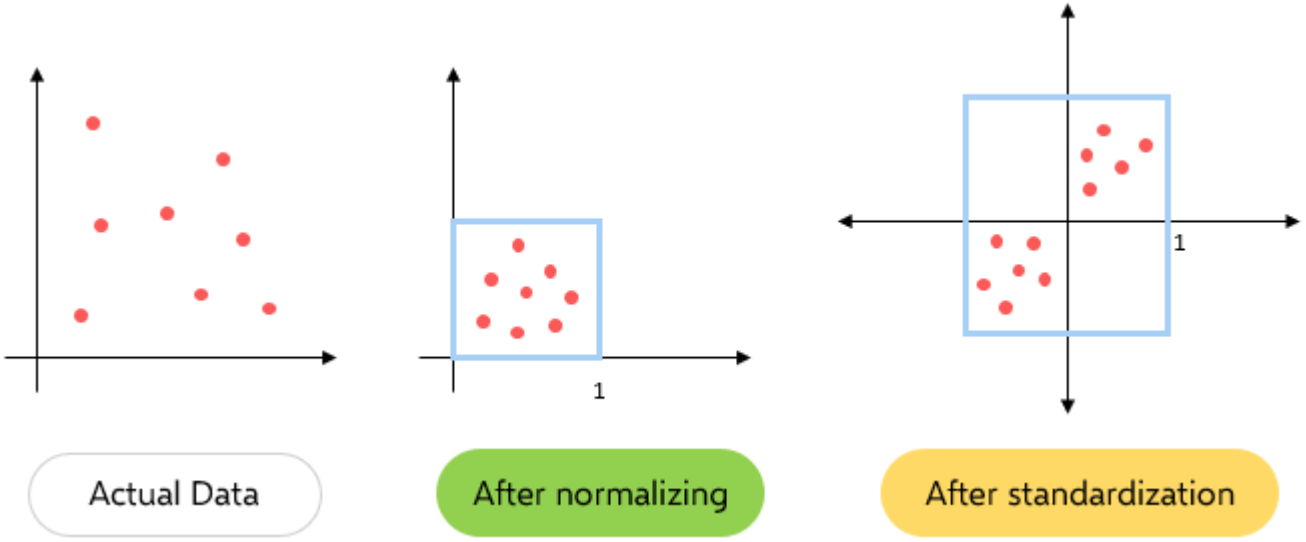
Normalização

- Analisar dados na mesma escala
  - 100 == 1 ? Sim, se considerarmos 100cm e 1 Metro
    - 100cm == 100cm ou 1m == 1m
  - Importante para algoritmos que analisam a distribuição espacial das características
  - A normalização é por atributo (características)

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal
297	59	1	0	164	176	1	0	90	0	1.0	1	2	1
243	57	1	0	152	274	0	1	88	1	1.2	1	1	3
269	56	1	0	130	283	1	0	103	1	1.6	0	0	3
215	43	0	0	132	341	1	0	136	1	3.0	1	0	3
83	52	1	3	152	298	1	1	178	0	1.2	1	0	3
152	64	1	3	170	227	0	0	155	0	0.6	1	0	3

Normalização

- MinMax (sklearn.preprocessing.MinMaxScaler) $\frac{x_i - \min(x)}{\max(x) - \min(x)}$
- Std (sklearn.preprocessing.StandardScaler) $\frac{x_i - \text{mean}(x)}{\text{stdev}(x)}$



Normalização

- MinMax (sklearn.preprocessing.MinMaxScaler) $\frac{x_i - \min(x)}{\max(x) - \min(x)}$

```
  Original: [ 171  216 -188  123  -61 -187   20 -161   33   -9]
Normalized: [[0.889 1.    0.    0.77  0.314 0.002 0.515 0.067 0.547 0.443]]
----------------------
  Original: [ 37.038  33.629   6.81  -13.202 -13.249 -44.837  -7.96   20.623  11.133
   8.058]
Normalized: [[1.    0.958 0.631 0.386 0.386 0.    0.45  0.8   0.684 0.646]]
```

- Standard (sklearn.preprocessing.StandardScaler) $\frac{x_i - \text{mean}(x)}{\text{stdev}(x)}$

```
  Original: [ 171  216 -188  123  -61 -187   20 -161   33   -9]
Normalized: [[ 1.264  1.588 -1.324  0.918 -0.409 -1.317  0.175 -1.13   0.269 -0.034]]
----------------------
  Original: [ 37.038  33.629   6.81  -13.202 -13.249 -44.837  -7.96   20.623  11.133
   8.058]
Normalized: [[ 1.425  1.278  0.129 -0.729 -0.731 -2.085 -0.504  0.721  0.314  0.182]]
```

Normalização

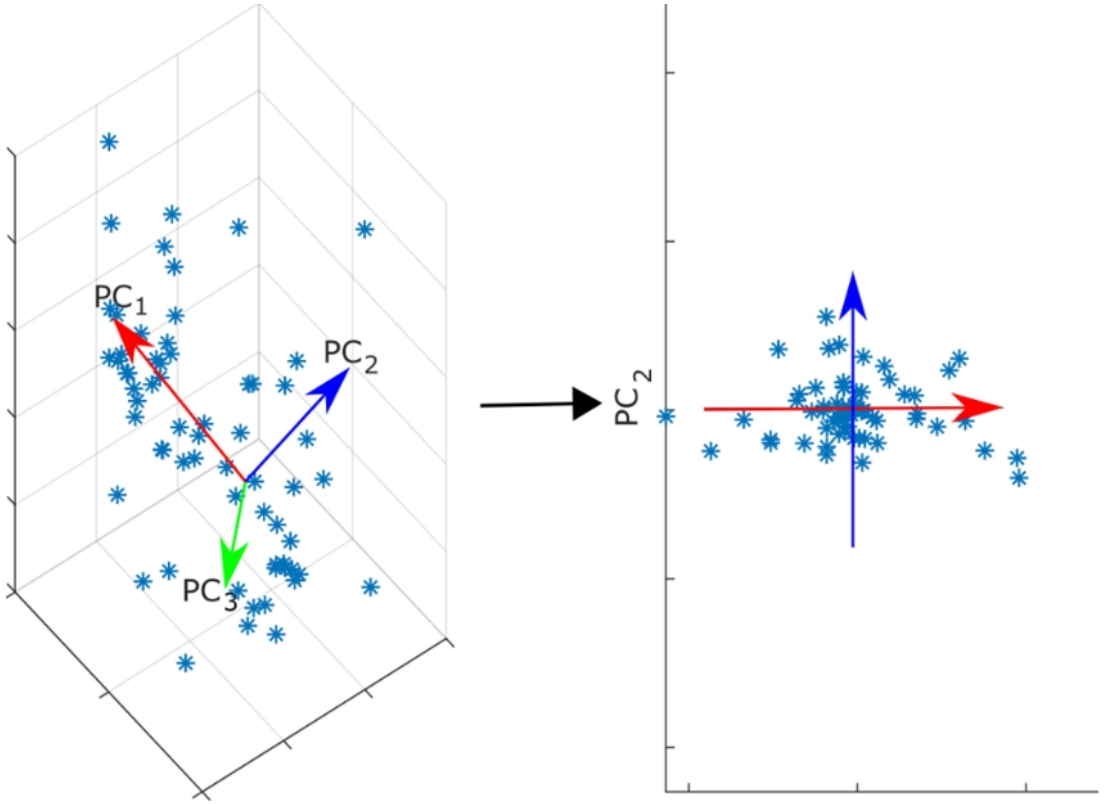
- E qual o impacto disso nos modelos?
  - KNN vs Naive Bayes

```
KNN  - Sem Normalização - ACC:  0.7222222222222222
              precision    recall  f1-score   support

           0       0.80      0.84      0.82        19
           1       0.77      0.77      0.77        22
           2       0.50      0.46      0.48        13

    accuracy                           0.72        54
   macro avg       0.69      0.69      0.69        54
weighted avg       0.72      0.72      0.72        54

KNN  - Com Normalização - ACC:  1.0
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        19
           1       1.00      1.00      1.00        22
           2       1.00      1.00      1.00        13

    accuracy                           1.00        54
   macro avg       1.00      1.00      1.00        54
weighted avg       1.00      1.00      1.00        54
```

```
Naive Bayes  - Sem Normalização - ACC:  0.944444444444
              precision    recall  f1-score   support

           0       0.90      1.00      0.95        19
           1       1.00      0.86      0.93        22
           2       0.93      1.00      0.96        13

    accuracy                           0.94        54
   macro avg       0.94      0.95      0.95        54
weighted avg       0.95      0.94      0.94        54

Naive Bayes  - Com Normalização - ACC:  0.944444444444
              precision    recall  f1-score   support

           0       0.90      1.00      0.95        19
           1       1.00      0.86      0.93        22
           2       0.93      1.00      0.96        13

    accuracy                           0.94        54
   macro avg       0.94      0.95      0.95        54
weighted avg       0.95      0.94      0.94        54
```

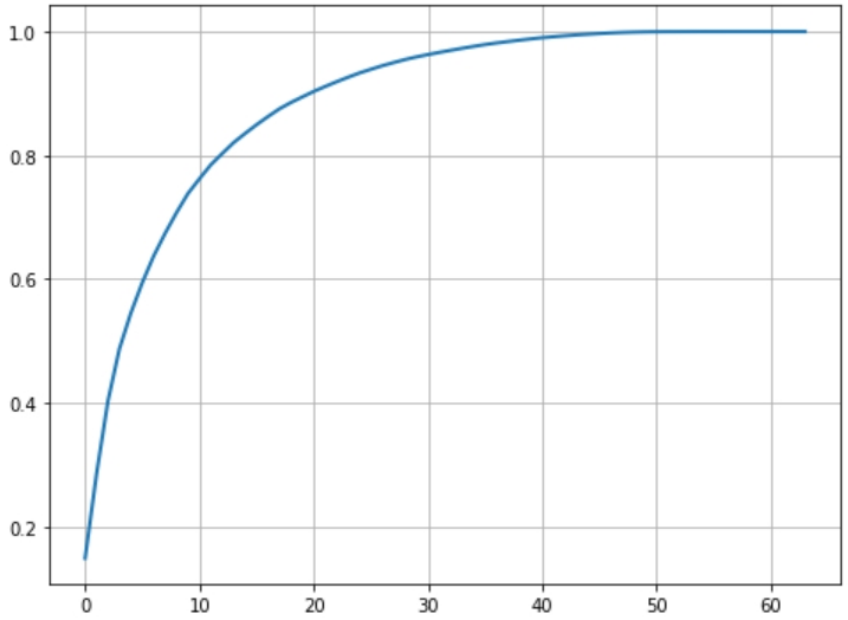
Redução de Características

- Redução de complexidade
- Ganho de Performance (tempo vs acc)
- Evitar o overfitting
- Seleção de features relevantes
  - Redução de ruído
  - Correlação
- Outras applicações
  - Visualização dos Dados
  - Compressão dos Dados



Redução de Características

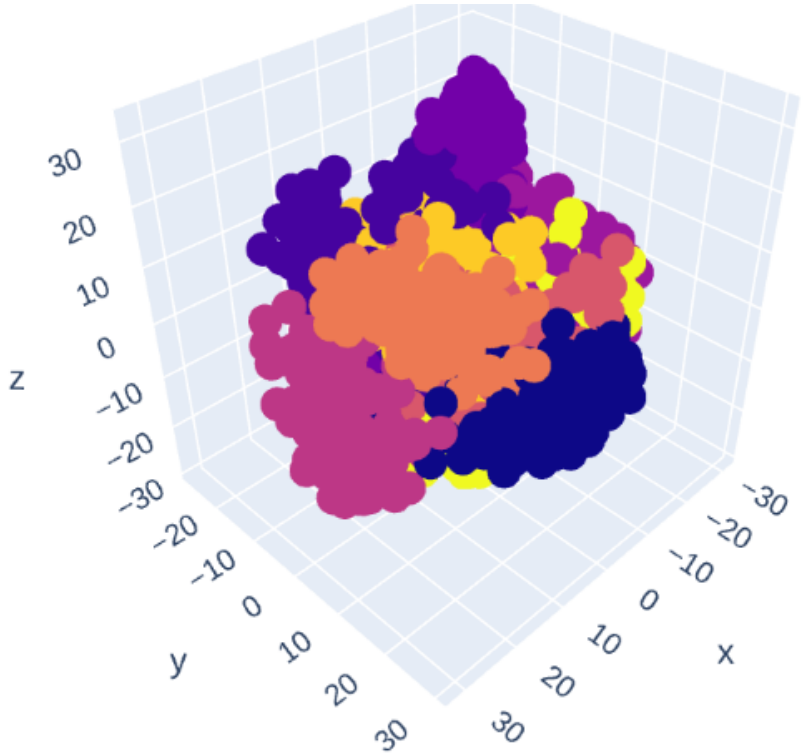
- PCA (sklearn.decomposition.PCA)
  - Autovalores e autovetores
  - Covariância e Correlação




```
n_comp: 2, acc: 0.5574
n_comp: 6, acc: 0.8333
n_comp: 10, acc: 0.820
n_comp: 14, acc: 0.827
n_comp: 18, acc: 0.833
n_comp: 22, acc: 0.838
n_comp: 26, acc: 0.831
n_comp: 30, acc: 0.833
n_comp: 34, acc: 0.814
n_comp: 38, acc: 0.822
n_comp: 42, acc: 0.820
n_comp: 46, acc: 0.827
n_comp: 50, acc: 0.811
n_comp: 54, acc: 0.824
n_comp: 58, acc: 0.820
n_comp: 62, acc: 0.838
```

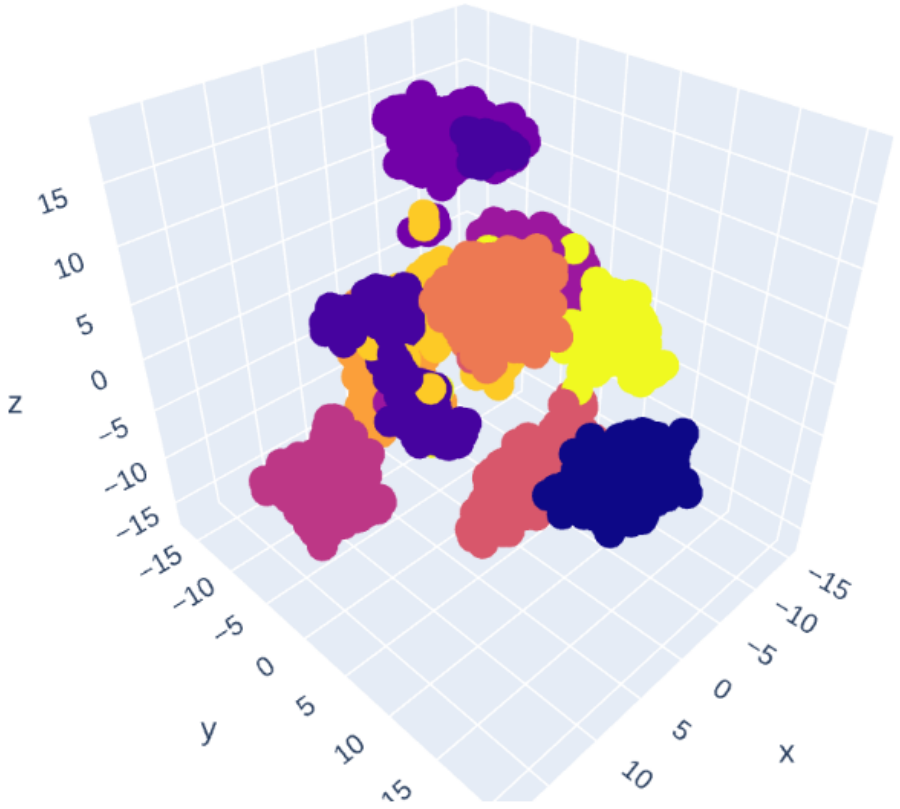
Visualização dos Dados

- PCA (n_comp=2 ou 3)



Visualização dos Dados

- TSNE (n_comp=2 ou 3)
  - Aprendizado de Máquina
  - Preserva Relações
  - Probabilidade em Alta e Baixa dimensão



Let’s code

- [LINK] Tópico_02_Aprendizado_Supervisionado_Normalizacao&Reducao.ipynb

- Off-topic:
  - Análise por Quartil
  - PCA para Redução e Compressão dos dados

Considerações Finais

- Normalização:
  - Vantagens
    - Ajuda a garantir que as variáveis de entrada tenham a mesma escala e * unidade de medida
    - Evita que valores extremos em uma variável dominem os valores de outras variáveis
    - Pode melhorar a precisão dos modelos de aprendizado de máquina
  - Desvantagens
    - Pode levar a uma perda de informação
    - Pode afetar a distribuição dos dados

Considerações Finais

- Redução de dimensionalidade:
  - Vantagens
    - Ajuda a lidar com problemas de alta dimensionalidade
    - Identifica as variáveis mais importantes para a previsão
    - Reduz a complexidade dos dados
  - Desvantagens
    - Pode levar a uma perda de informação
    - Pode afetar a interpretabilidade dos dados